SYNTHESIS OF 3D VIRTUAL AUDITORY SPACE VIA A SPATIAL FEATURE EXTRACTION AND REGULARIZATION MODEL[1]

Jiashu Chen, Barry D. Van Veen, and Kurt E. Hecox

University of Wisconsin-Madison
600 Highland Ave. Rm. H6/573, Madison, WI 53792, USA

1 Introduction

Measured head-related transfer functions (HRTFs) are increasingly employed to synthesize a "realistic" 3D auditory display over earphones. It has been reported that sounds processed through HRTFs are perceived as if they are from the corresponding HRTF location and distance [12, 10]. These findings have led to new developments in sound localization research and applications related to acoustical virtual environments.

The HRTFs are usually derived from measured free-field and ear canal recordings of a broadband stimulus at specified locations. Recordings must be made for each virtual location to be simulated. Simulation of the entire 3D space on a fine grid thus requires a very large number of recordings. Obtaining a very large number of recordings is often impractical because of time and data storage limitations. When multiple sound sources or room acoustics are employed to simulate complex acoustical environments the number of HRTFs involved can easily exceed the real time processing capability of present generation computers. Furthermore, even if these recordings are made, they only represent a discrete sampling of the auditory space.

A functional representation for the HRTFs is desirable as it overcomes many of the limitations associated with use of measured HRTFs. For example, it provides a continuous representation of auditory space such that the synthesis of HRTF at any given spatial location can be performed by functional evaluation of the model. In this paper we propose such a model that establishes a mathematical representation of the external ear transformation characteristics based on spatial feature extraction and regularization.

2 The HRTFs

Free-field and ear canal recordings are collected in semi-anechoic room using the experimental set up reported in [8]. Ear canal recordings are collected using B&K 1/2 in. microphones mounted on the acoustic couplers in a KEMAR head. The sound stimulus is a 20-μs impulse delivered to a Radio-Shack middle range speaker. The speaker is moved on a sphere in 4.5° intervals both in elevation and azimuth by a computer-controlled servo system. The KEMAR head is at the center of this sphere at a distance of 75 cm from the speaker. The frontal direction is defined as zero degrees elevation and azimuth. The vertex is 90

[1]Raw data provided by Dept. of Neurophysiology, University of Wisconsin-Madison. This work was supported by NIH grant R01 NS1 6436.

0-7803-1363-1/93 $3.00 © 1993 IEEE

degrees elevation. Directions corresponding to left and right ears are defined as -180 and +180 degrees of azimuth, respectively. Figure 1 illustrates this coordinate system. For each elevation 80 different azimuthal measurements are recorded. There are 29 elevation positions from -36 degrees of elevation to 90 degrees, giving a total of 2320 virtual positions. Because of mechanical limitations only 2188 sets of recordings are made. The missing measurements are at the lower back region. The HRTFs are derived from the recordings using the spectral analysis method of system identification [6] to reduce the effects of noise and artifacts [2].

3 The spatial feature extraction and regularization model of measured HRTFs

Let $\mathbf{h}_j, j = 1, ..., P$ be an N-by-1 complex-valued vector, representing the jth HRTF from a set of P measured HRTFs. Exploiting the observation that all the HRTFs have some degree of similarity, we hypothesize that the HRTFs are "clustered" in a subspace of the N-dimensional space. Let the dimension of this subspace be M. If $M \ll N$, then an efficient low-dimensional representation of the HRTFs is possible.

Assume $\mathbf{h}_j$ is represented without error by an expansion of the form

$$\mathbf{h}_j = \mathbf{h}_{dj} + \mathbf{h}_{av} = \sum_{i=1}^{N} w_{ij}\mathbf{q}_i + \mathbf{h}_{av} = \mathbf{Q}\mathbf{w}_j + \mathbf{h}_{av}, \tag{1}$$

where the columns of $\mathbf{Q} = [\mathbf{q}_1 \ \ldots \ \mathbf{q}_N]$ are a linearly independent basis for N dimensional space and $\mathbf{w}_j = [w_{1j} \ \ldots \ w_{Nj}]^T$ is a vector representing the coordinates of $\mathbf{h}_{dj}$ with respect to the basis $\mathbf{Q}$. The vector $\mathbf{h}_{av}$ is the sample average , defined as $\mathbf{h}_{av} = \frac{1}{P}\sum_{j=1}^{P}\mathbf{h}_j$, and $\mathbf{h}_{dj}$ is the HRTF with the sample-average removed. Let the columns of $\mathbf{Q}$ be an orthonormal basis, that is,

$$\mathbf{q}_i^H\mathbf{q}_k = \begin{cases} 1 & i = k \\ 0 & i \neq k \end{cases} \tag{2}$$

When the orthonormality condition is satisfied, the weight vector $\mathbf{w}_j$ is given by

$$\mathbf{w}_j = \mathbf{Q}^H\mathbf{h}_{dj} \tag{3}$$

or, alternatively by

$$w_{ij} = \mathbf{q}_i^H\mathbf{h}_{dj}, \qquad i = 1, ..., N. \tag{4}$$

Therefore, $\mathbf{w}_j$ is simply an orthonormal transformation of the vector $\mathbf{h}_{dj}$. Our goal is to find an orthogonal transformation matrix $\mathbf{Q}$ such that only a subset

$$\{w_{1j}, w_{2j}, ..., w_{Mj}\}$$

of $\mathbf{w}_j$ are needed to accurately represent $\mathbf{h}_{dj}$, while the remaining $N - M$ components of $\mathbf{w}_j$ are approximately zero for all $\mathbf{h}_{dj}$. It can be shown [2] that $\mathbf{Q}$ is obtained by solving the following eigen problem

$$\mathbf{R}_h\mathbf{Q} = \mathbf{\Lambda}\mathbf{Q} \tag{5}$$

where $\mathbf{R}_h = \frac{1}{P}\sum_{j=1}^{P} \mathbf{h}_{dj}\mathbf{h}_{dj}^H$ is the sample covariance matrix of the $\mathbf{h}_{dj}$, $\mathbf{\Lambda} = diag\{\lambda_1, ..., \lambda_N\}$ is a diagonal matrix formed from the N eigenvalues of $\mathbf{R}_h$. Note that $\lambda_i, i = 1, ..., N$ represent the sample variance of $\mathbf{h}_{dj}$ projected onto eigenvector $\mathbf{q}_i, i = 1, ..., N$. Hence, the relative size of the eigenvalues determines which M $\mathbf{q}_i$ to choose. The M eigenvectors corresponding to M largest eigenvalues are the optimal choices for (1) in the mean square error sense [2].

Recall that there are 80 HRTFs sampled on each elevation circle. The elevation circles at higher (and lower) elevations span much smaller solid angles than those at or close to the equator. To prevent the higher elevation HRTFs from dominating the sample covariance matrix, we form sub-covariance matrices constructed from the 80 HRTFs with the same elevation. The sample covariance matrix $\mathbf{R}_h$ is then formed as a weighted sum of the subcovariance matrices with the weighting proportional to the solid angle spanned by the corresponding elevation circle. For the KEMAR data $M = 16$ eigenvalues represent approximately 99.9% of the energy in the measured HRTFs.

Once the M $\mathbf{q}_i$ are chosen, any $\mathbf{h}_{dj}$ is approximated as a linear combination of $\mathbf{q}_i$'s, that is,

$$\mathbf{h}_{dj} = \sum_{i=1}^{M} w_{ij}\mathbf{q}_i. \tag{6}$$

where the $w_{ij}, i = 1, ..., M$ are determined by projecting $\mathbf{h}_{dj}$ onto $\mathbf{q}_i, i = 1, ..., M$ as shown in (4).

The expansion defined in (1) is thus in terms of eigenvectors of the covariance matrix $\mathbf{R}_h$. This expansion is known the *Karhunen-Loève expansion* (KLE) and the transformation $\mathbf{w}_j = \mathbf{Q}^H\mathbf{h}_{dj}$ is termed the *Karhunen-Loève transform* (KLT) [1].

It is clear that (6) implements a low-dimensional representation of all $\mathbf{h}_{dj}$. Since the elements of $\mathbf{h}_{dj}$ are DFT coefficients, the $\mathbf{q}_i$ have a physical interpretation as *eigen-transfer functions* (EFs). They represent the frequency characteristics of the auditory space described by the measured HRTFs. The weights $w_{ij}, i = 1, ..., M; j = 1, ..., P$ represent the spatial characteristics of the auditory space since they describe the spatial attributes of the measured HRTFs. We assume the $w_{ij}, i = 1, ..., M, j = 1, ..., P$ are samples of underlying continuous functions $w_i(\theta, \phi), i = 1, ..., M$, termed *spatial characteristic functions* (SCFs). Substituting the continuous SCFs into (1) and setting the summation upper limit to M, we obtain the functional representation for the HRTF

$$\mathbf{h}(\theta, \phi) = \sum_{i=1}^{M} w_i(\theta, \phi)\mathbf{q}_i + \mathbf{h}_{av}. \tag{7}$$

A thin-plate spline model [9] is employed to regularize the samples of the SCFs derived from the measured HRTFs. "Regularize" refers to reconstruction of the SCFs $w_i(\theta, \phi)$ from the P noise contaminated samples $w_{ij}, j = 1, ..., P$. Regularization results in a functional representation for each SCF and thus enables interpolation of the SCF between sample points. It is accomplished by solving the spline-based regularization problem [9]:

$$\min_{\hat{w}_i(\theta,\phi)} \sum_j \sum_k (w_i(\theta_j, \phi_k) - \hat{w}_i(\theta_j, \phi_k))^2 + \lambda \|S(\hat{w}_i(\theta, \phi))\|^2, \tag{8}$$

where $\{w_i(\theta_j, \phi_k), j = 1, ..., J, k = 1, ..., K\}$ are the samples of ith SCF, $\hat{w}_i(\theta, \phi)$ is the functional approximation to the ith SCF, and λ is the regularization parameter. Viewing θ and ϕ as coordinates in a two dimensional rectangular coordinate system,

$$\|S(\hat{w}(\theta,\phi))\|^2 = \int_\Phi \int_\Theta d\theta d\phi \{[\frac{\partial^2 \hat{w}(\theta,\phi)}{\partial\theta^2}]^2 + 2[\frac{\partial^2 \hat{w}(\theta,\phi)}{\partial\theta\partial\phi}]^2 + [\frac{\partial^2 \hat{w}(\theta,\phi)}{\partial\phi^2}]^2\}. \quad (9)$$

The regularization parameter λ controls the trade-off between the smoothness of the solution and the fidelity to the measured data. The value of λ is determined via generalized cross validation [9]. In practice, the regularization algorithm is implemented using a publicly available software package Rkpack [4].

The descriptive term *Spatial Feature Extraction and Regularization* (SFER) model is adopted because the KLE is used to extract spatial features (SCF samples) from the HRTFs which are then regularized using the generalized spline model.

4 Results

The performance of the model is acoustically validated using a large number of measured HRTFs sampled over the upper 3/4 sphere for the right ear of a KEMAR manikin. We have shown that 16 EFs are sufficient to represent 99.9% of the energy in the measured HRTFs. SCF samples on a 9-degree grid (azimuth and elevation) are used to determine the model parameters. The approximation error and predictive power of the model are examined by comparing the modeled and measured HRTFs on a on a 4.5 degree grid over a total of 2188 locations. In the frontal and ipsi-lateral areas the errors are on the order of hundredths of one percent. The errors are larger in the back and contralateral areas because of reduced HRTF amplitude. Figure 2 depicts several measured and modeled HRTFs on the horizontal plane. Figure 3 depicts several comparisons on the median plane. Figure 4 depicts HRTFs interpolated on a 1.5-degree grid along the median plane as a mesh plot. Interesting relationships between the SCFs and the external ear geometry have also been observed [2].

5 Summary

Models for the HRTF based on principal component analysis methods are reported in [7, 5]. However, these approaches only model the amplitude spectrum and represent up to 92% of the energy in the measured HRTF amplitude. Furthermore, these methods are not based on functional approximations of measured SCFs and thus are not able to interpolate the HRTF between measured directions. The beamforming model [3] is limited to modeling small sectors of the auditory space for computational reasons. The SFER model described here is ideally suited for efficient simulation of virtual auditory space. The identified eigen transfer functions and extracted spatial features provide a set of parameters potentially useful for categorizing the transformation characteristics among different subjects [2]. This model establishes a framework for study and physical interpretation of HRTFs.

References

[1] N. Ahmed and K. R. Rao. *Orthogonal transforms for digital signal processing.* Springer-Verlag, New York, 1975.

[2] J. Chen. *Auditory Space Modeling and Virtual Auditory Environment Simulation.* PhD thesis, University of Wisconsin-Madison, H6/573, 600 Highland Ave., Madison, WI 53792, 1992.

[3] J. Chen, B. D. Van Veen, and K. E. Hecox. External ear transfer function modelling: a beamforming approach. *J. Acoust. Soc. Am.*, 92(4):1933-1944, 10. 1992.

[4] C. Gu. Rkpack and its applications: Fitting smoothing spline models. Technical Report 857, Department of Statistics, University of Wisconsin-Madison, 1989.

[5] D. J. Kistler and F. L. Wightman. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *J. Acoust. Soc. Am.*, 91:1637–1647, 3 1992.

[6] L. Ljung. *System Identification: Theory for the User.* prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632, 1987.

[7] W. L. Martens. Principal components analysis and resynthesis of spectral cues to perceived direction. *The International Computer Music Conference*, pp. 274–281, San Francisco, CA, 1987. International Computer Music Assiociation. J. beauchamp, editor.

[8] A. D. Musicant, J. C. Chan, and J. E. Hind. Direction-dependent spectral properties of cat external ear: New data and cross-species comparisons. *J. Acoust. Soc. Am.*, 87(2):757–781, 2 1990.

[9] G. Wahba. *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1990.

[10] E. M. Wenzel. Issues in the development of virtual acoustic environments. *J. Acoust. Soc. Am.*, 92(2):2332, 1992.

[11] E. M. Wenzel, F. L. Wightman, and S. H. Foster. A virtual display system for conveying three-dimensional acoustic information. *Proc. Hum. Fac. Soc.*, 32:86–90, 1988.

[12] F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening: I: Stimulus synthesis. *J. Acoust. Soc. Am.*, 85(2):858–867, 1989.

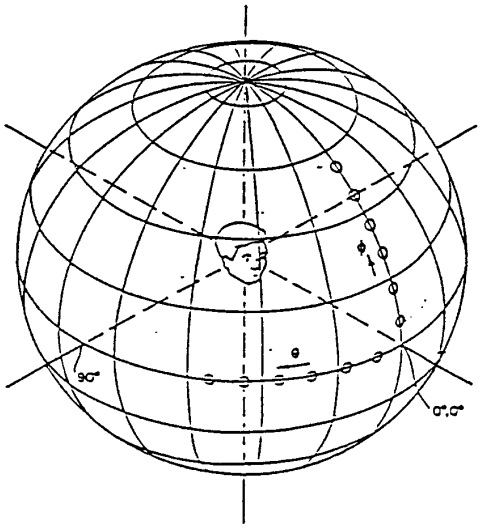

Figure 1: Data acquisition coordinate system

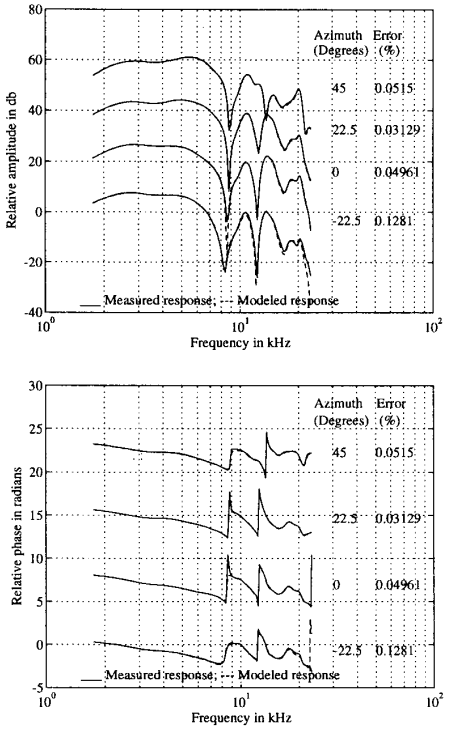

Figure 2: Comparisons between measured and modeled HRTFs at several locations on the horizontal plane

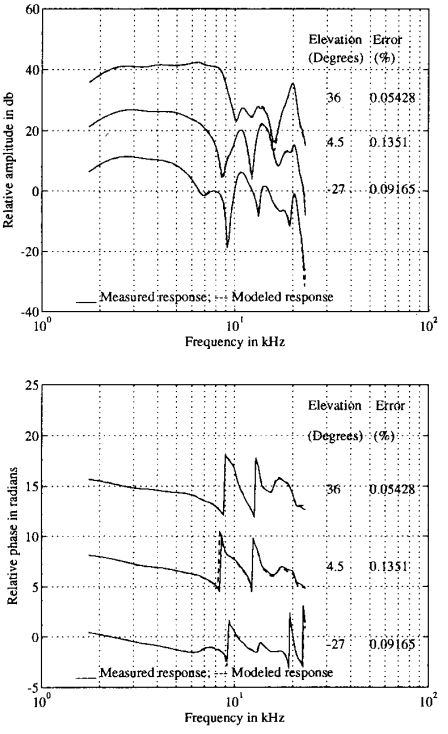

Figure 3: Comparisons between measured and modeled HRTFs at several locations on the median plane

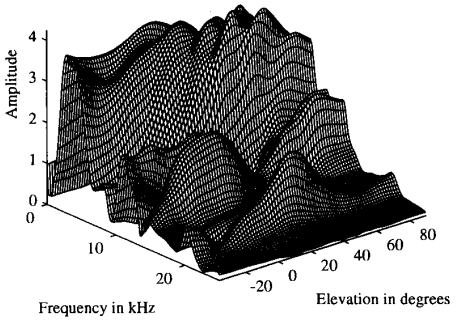

Figure 4: Mesh plot of the model based HRTF amplitudes along the median plane. ($\theta = 0°$), ϕ is from $-36°$ to 90o at 1.5° intervals.